International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 4, Issue 2, Apr 2014, 253-258

'ANUVADAK' ENGLISH TO HINDI TEXT TRANSLATOR

SUDHIR D BAGUL[1] & UDAY B JOSHI[2]

[1]ME Student, Department of Computer Engineering, K J Somaiya College of Engineering, Vidyavihar (East), Mumbai, India

[2]Associate Professor, Department of Computer Engineering, K J Somaiya College of Engineering, Vidyavihar (East), Mumbai, India

ABSTRACT

Machine based translation, sometimes referred to by the abbreviation MT, is a sub-field of computational linguistics that investigates the use of software to translate text from one natural language to another. As more number of people are using computers for their day to day work the need of good translator is been arrived, to fulfill that need many translators are proposed and implemented. Translating from one language to another needs a detailed knowledge of languages to get meaningful translation. Available approaches for machine based translations are appropriate but may fail when language's nature and grammar rule differs, in such cases, possibilities are more to get meaningless translation. In this paper we propose a new approach called as hybrid approach by combining Rule based and Example based translation methods. This will help to overcome some of the issues and errors in existing translators. In this method we identify the type of sentence and then translate it from English to Hindi.

General Terms: Translator, Source Language, Target Language, Translation, Machine Based Translation

KEYWORDS: Natural Language Processing (NLP), Part of Speech (POS), Parser, Rule Based Approach, Example Based Approach, Source Language, Target Language

INTRODUCTION

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at once or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications [1].

Natural Language Interface gives the user freedom to interact with the computer in a natural language like English, Hindi, Marathi or any other language used for day to day communication. One of the important goals of applied computational linguistics is a fully automatic machine translation between such natural languages. This is important because communication between people from different linguistic backgrounds still poses as a major problem [2] [3]

It is important to understand the differences in the syntactic structures of source and target languages in multi-lingual machine translators. Most Indian languages are derivatives of Sanskrit which is a historical Indo-Aryan language with extremely rich grammar. 'Hindi', which is considered as the target language in the design of Anuvadak, is an important derivative language of Sanskrit []

TRANSLATION PROCESS

The Human Translation Process may be described as:

- Decoding the meaning of the source text; and
- Re-encoding this meaning in the target language.

Behind this ostensibly simple procedure lies a complex cognitive operation. To decode the meaning of the source text in its entirety, the translator must interpret and analyze all the features of the text, a process that requires in-depth knowledge of the grammar, semantics, syntax, idioms, etc., of the source language, as well as the culture of its speakers. The translator needs the same in-depth knowledge to re-encode the meaning in the target language.

Therein lies the challenge in machine translation: how to program a computer that will "understand" a text as a person does, and that will "create" a new text in the target language that "sounds" as if it has been written by a person.

This problem may be approached in a number of ways
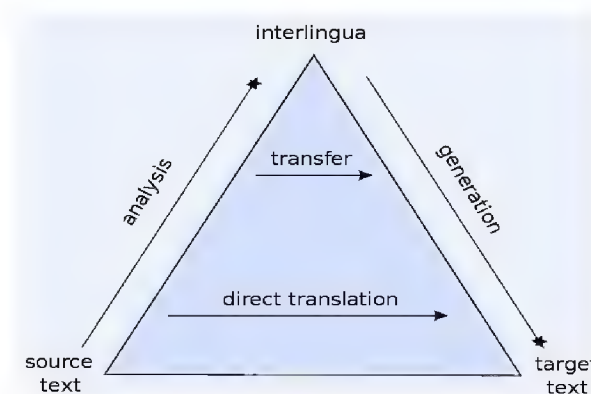
Figure1: Translation Process: Courtesy Wikipedia

Many available translators fails in English to Hindi translation because they do not identify the type of sentence in both source and target languages appropriately and hence results in some meaningless or unwanted translation.

PROPOSED DESIGN OF ANUVADAK

To overcome and solve some of the issues like gender and tense recognition in English to Hindi translation Anuvadak design is proposed as follows; which first identifies the type of sentence from source language which user wants to get translated and then will translate the sentence in target language.
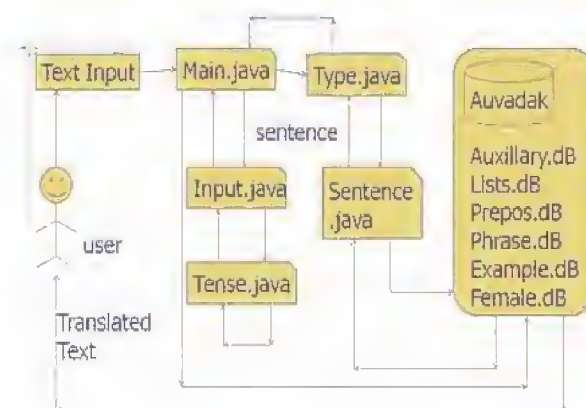
Figure 2: 'Anuvadak' Design

As shown in above architecture the translator named as 'Anuvadak' accepts an input from user in the form of English text then following steps are followed:

Step 1: Identify the Type of Sentence given by user

Step 2: Identify the Gender, Phrases, Prepositions and Auxiliaries from sentence

Step 3: Identify Tense of sentence

Step 4: Give "Hindi" translated output to user.

Implementation of Anuvadak English to Hindi Translation Parser Issues

There are various problems faced by a parser if not endowed with intelligence. It requires intelligence to address different issues like:

- Removal of inanimate gender in English and conversion to one of the genders available in Hindi. For instance, 'chair' – an inanimate gender in English when translated to Hindi gives कुर्सी 'kursi'-feminine gender.
- After the pre-processing of English sentences, the POStagged English words are obtained. These words are translated to their Hindi counterparts at the word level. This can be done with the help of an English-to-Hindi dictionary. The Hindi words thus obtained, need to be organized in the SVO form to get the complete translated sentence. This involves the addition of cases (called karaka) in the right places to maintain correct syntactic structure in Hindi. In Hindi grammar, cases are suffixes or postpositions that directly connect to verbs. Cases are prefixes only for vocative ('Sambodhan') cases.

  Consider the Example

  "Ram hit Shaam." It is translated to,
  राम ने शाम को पीटा |
  (Ram ne Shaam ko peeta)

  Addition of ने and को is necessary to maintain the syntactic structure of the translation
- If a word in English has various meaning, then deduce the correct meaning using POS tagging followed by word sense disambiguation Example: "Tarun mail the mail." Here both occurrences of word 'mail' mean differently. This ambiguity is solved using POS tagging. The grammar of the tokens can be identified and classified; that will help identify that the first occurrence of 'mail' refers to a verb whereas, the second occurrence is a noun.
- Some sentences are considered implicitly ambiguous if the word meaning is evaluated only at the current sentence level. Example: "I hate annoying teachers." In such cases the only solution to infer the correct sense of the ambiguous word(s) in the sentence is to consider the context. This can be done by taking into account an acceptable context window of size 'n'.
- Some statements when translated into Hindi require an implied subject.

  Example: "Catch the ball!"

  Here the Subject ('you') is implied and understood.

Some of above mentioned issues get resolved by adapting the 'Anuvadak' approach.

ANUVADAK IS A HYBRID METHOD

There are many approaches are used to get machine based translation two of them are Rule Based and Example Based.

Rule-Based

The rule-based machine translation paradigm includes transfer-based machine translation, interlingual machine translation and dictionary-based machine translation paradigms [2]

Example-Based

Example-based machine translation (EBMT) approach was proposed by Makoto Nagao in 1984. It is often characterised by its use of a bilingual corpus as its main knowledge base, at run-time. It is essentially a translation by analogy and can be viewed as an implementation of case-based reasoning approach of machine learning. [2]

As we have proposed in our system we check the rules of sentence types and we also save the history in database to get sentence record we use both rule based and example based approaches in 'Anuvadak' design hence it is a hybrid approach of translation.

IMPLEMENTATION OF ANUVADAK

In implementation we recommend to use JAVA as front end and MySQL as a back end to store the examples and rules of parsing method. Some part is implemented as follows:
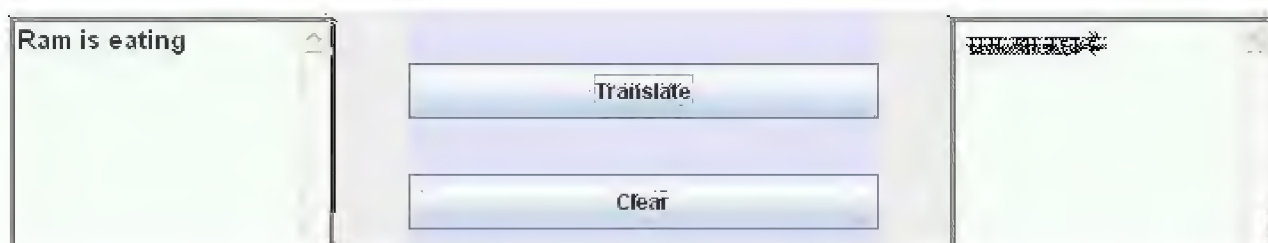
Figure 3: Anuvadak Implementation

In above figure implemented parser indentifies the first type of sentence i.e. simple sentence and translates it in Hindi.

The Anuvadak Performs the Following Two Functions

- Word Translation
- Sentence Translation
- **Word Translation**

  Steps: Start.

  The user inputs an English word for translation. The input string is processed and validated.

The word is searched in the English-Hindi word database.

If the word is found, it is returned as output. The output is displayed to the user.

Stop.

- **Sentence Translation Steps**

  The Translation of Sentence Includes Two Parts:

  - Finding example sentence based on POS.
  - Translation using Example based algorithm by the following functions: -
    - No Change
    - Word Addition
    - Word Replacement
    - Word Deletion
  - Tense Recognition
  - Gender Recognition
  - Singularity Recognition
  - Phrase Recognition

Finding Example Sentence Based on POS

Database has an additional column called Id which will be given according to Part of Speech (POS). POS function will give us the following details:

- Subject
- Auxiliary Verb
- Main Verb
- At most two Objects and their Adjectives
- Connecting Preposition

Taking Permutation and Combination of the above Following

POS we have about nine types of sentences.

Subject+ Auxiliary Verb Main Verb Subject+

Auxiliary Verb Main Verb+Object1

Subject+Auxiliary Verb Main

Verb+Adjective1+Object1 Subject+Auxiliary Verb

Main Verb+Preposition+Object2 Subject+Auxiliary

Verb Main Verb+Preposition+Adjective2+Object2

Subject+ Auxiliary Verb Main

Verb+Object1+Preposition+Object2

Subject+Auxiliary Verb+Main Verb+Object1+Preposition+Adjective2+Object2

Subject+Auxiliary+Verb Main

Verb+Adjective1+Object1+Preposition+Object2 Subject+Auxiliary Verb Main

Verb+Object1+Adjective1+Preposition+Adjective2+Objective2

ACKNOWLEDGMENTS

Our thanks to K J Somaiya College of Engineering, Vidyavihar, Mumbai for their support and making available all needed resources.

REFERENCES

1. Liddy, E. D. In Encyclopedia of Library and Information Science, 2nd Ed. Marcel Decker, Inc.
2. Rekha S. Sugandhi, RitikaShekhar, TarunAgarwal, Rajneesh K. Bedi, Vijay M. Wadhai 978-1-4673-0126-8/11/$26.00c 2011 IEEE, “Issues in Parsing for Machine Aided Translation from English to Hindi”
3. S Samantary “Text Mining approach for resolving cases of multiple parsing in machine aided translation of Indian Languages” in proceedings of 4th IEEE conference, April 2007.
4. GB THEORY BASED HINDI TO ENGLISH TRANSLATION SYSTEM by Alka Choudry & Manjeet Singh 978-1-4244-4520-2/09/$25.00 ©2009 IEEE.
5. Text Studies towards Multi-lingual Content Mining for Web Communication by Kolla Bhanu Prakash, M. A. Dorai Rangaswamy, Arun Raja Raman 978-1-4244-9008-0/10/$26.00 ©2010 IEEE
6. White paper on Design and Development of Translator's Workbench for English to Indian Languages by Akshi Kumar
7. White paper on English-Hindi Automatic Word Alignment with Scarce Resources by Deepa Gupta & Eknath Venkataramani